MTSplice predicts effects of genetic variants on tissue-specific splicing

Anonymous Authors[1]

Abstract

Tissue-specific splicing of exons plays an important role in determining tissue identity. However, computational tools predicting tissue-specific effects of variants on splicing are lacking. To address this issue, we developed MTSplice (Multi-tissue MMSplice), a neural network which quantitatively predicts effects of human genetic variants on splicing of cassette exons in 56 tissues. MTSplice combines the state-of-the-art predictor MMSplice, which models constitutive regulatory sequences, with a new neural network which models tissue-specific regulatory sequences. MTSplice outperforms MMSplice on predicting effects associated with naturally occurring genetic variants in most tissues of the GTEx dataset. Furthermore, MTSplice predicts that autism-associated *de novo* mutations are enriched for variants affecting splicing specifically in the brain. We foresee MTSplice to be useful for functional prediction and prioritization of variants associated with tissues-specific disorders.

1. Introduction

Splicing defects account for an important fraction of the genetic basis of human diseases (López-Bigas et al., 2005). Some of these splicing defects are specific to disease-relevant tissues. For instance, individuals affected by autism spectrum disorder (ASD) frequently present mis-splicing of brain-specific exons (Parras et al., 2018) as well as an enrichment of *de novo* mutations in brain-specific exons (Uddin et al., 2014). Hence, computational tools that can predict the tissue-specific effects of genetic variants on splicing would be relevant for understanding the genetic basis of tissue-specific diseases such as ASD.

Many computational tools have been developed to predict splice sites or splicing strength from sequence (Yeo & Burge, 2004; Desmet et al., 2009; Jian et al., 2014; Rosenberg et al., 2015; Xiong et al., 2015; Cheng et al., 2019; Sonnenburg et al., 2007; Jaganathan et al., 2019). However, tools are lacking for predicting tissue-specific effects of genetic variants on splicing in human cells. Barash et al. developed the first sequence-based model predicting tissue-specific splicing in mouse cells (Barash et al., 2010). The model integrates regulatory sequence elements to qualitatively predict whether the inclusion of a cassette exon increases, decreases, or remains at a similar level from one tissue to another tissue. This model was further improved to predict directional changes between tissues along with discretized Ψ categories (Low, -Medium, and -High) within a tissue by using Bayesian neural network with hidden variables (Xiong et al., 2011). A similar Bayesian neural network (SPANR) was later on trained on human data (Xiong et al., 2015). However, SPANR was evaluated only for predicting the largest effect across all investigated tissues. Hence, the performance of SPANR on any given tissue is unclear. Moreover, the publicly available SPANR does not allow performing tissue-specific predictions.

Here, we developed MTSplice (Multi-tissue MMSplice), a model that predicts tissue-specific splicing effects of human genetic variants. MTSplice adjusts the tissue-agnostic state-of-the-art predictor MMSplice (Cheng et al., 2019) with the predictions of TSplice (Tissue-specific Splicing), a newly developed deep neural network predicting tissue-specific variations of Ψ from sequence and trained on 56 human tissues using multi-task learning. Performance of MTSplice is demonstrated by predicting tissue-specific variations of Ψ associated with naturally occurring genetic variants of the GTEx dataset as well as investigating brain-specific splicing effect predictions for autism-associated variants.

2. Results

To train a tissue-specific model of splicing, we considered the alternative splicing catalog of the transcriptome ASCOT (Ling et al., 2020). Because the ASCOT annotation and quantification pipeline is annotation-free, it also covers non-annotated exons. Altogether, ASCOT provides Ψ values for 61,823 cassette exons across 56 tissues including 53 tissues from the GTEx dataset (GTEx Consortium, 2013) and additional RNA-Seq data from peripheral retina. Of

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at the ICML 2020 Workshop on Computational Biology (WCB). Do not distribute.

note, these tissue-specific values are flagged as missing when the corresponding gene is not expressed (Ling et al., 2020).

2.1. TSplice predicts tissue-specific Ψ

We trained a new neural network, TSplice, to predict tissue-specific Ψ values from sequence and tissue-averaged Ψ (Methods). TSplice considers the 300 nt flanking either side of the exon and the first and last 100 nt of the exon body. TSplice is a convolutional neural network Figure1 in which positional effects of sequence elements relative to splice sites are modeled using spline transformations (Avsec et al., 2018). TSplice was trained on the ASCOT dataset (Ling et al., 2020). ASCOT includes de novo annotated exons provides Ψ values for 61,823 cassette exons across 56 tissues including 53 tissues from the GTEx dataset (GTEx Consortium, 2013) and additional RNA-Seq data from peripheral retina. We report our performances on three held-out chromosomes (chromosome 2, 3 and 5).

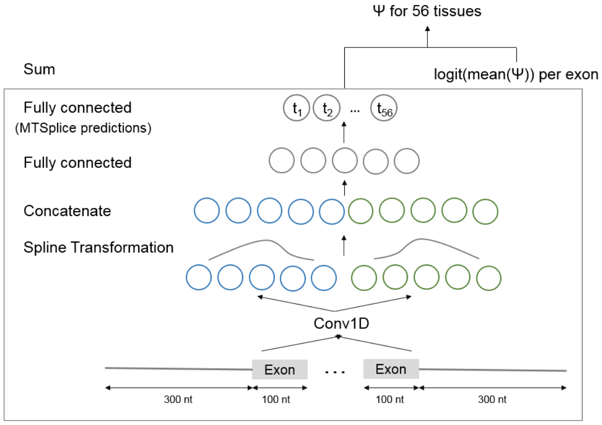

Figure 1. Model architecture to predict tissue-specific percent spliced-in. The model TSplice consists of one convolution layer with 64 length-9 filters capturing sequence elements. This is followed by two spline transformation layers modulating the effect of sequence elements depending on their position relative to the acceptor splice sites (leftmost layer) and the donor (rightmost layer). The outputs of the two spline transformation layers are concatenated and global average pooling is applied along the sequence dimension. This is then followed by feeding two consecutive fully-connected layers. The last fully-connected layer outputs a 56 dimension vector which are the predicted log odd ratios of tissue-specific Ψ versus tissue-averaged Ψ for the 56 tissues of the ASCOT dataset. Natural scale tissue-specific Ψ are obtained by adding predicted odd ratios with measured tissue-averaged Ψ. In total, the model has 8,024 trainable parameters.

The performance of TSplice was first assessed on test data by comparing the observed against the predicted log odd ratios of tissue-specific Ψ for 1,621 exons ("variable exons") with Ψ deviating from the tissue-averaged Ψ by at least 0.2 in at least one tissue and for which the gene is expressed in at least 10 tissues (Figure2 for Retina eye as an example, Spearman correlation = 0.27). The predictions positively correlated with the measurements in all tissues and showed a median Spearman correlation of 0.22 (Figure2B). The performance were higher for tissues of the central nervous system (Figure2C), possibly because central nervous system tissues harbor similar splicing patterns and because they are well represented in the ASCOT dataset.

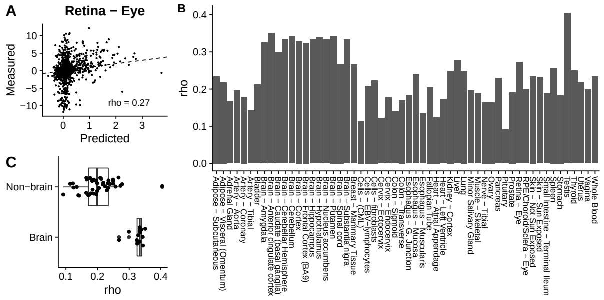

Figure 2. Evaluating TSplice on predicting tissue-associated differential splicing. **(A)** Predicted versus measured $\delta\,\mathrm{logit}(\Psi_{e,t})$ for the Retina-Eye tissue, representative of the typical performance of our model. **(B)** Spearman correlation between predicted and measured $\delta\,\mathrm{logit}(\Psi_{e,t})$ for all tissues. **(C)** Distribution of Spearman correlations between predicted and measured $\delta\,\mathrm{logit}(\Psi_{e,t})$ for brain tissues and non-brain tissues.

2.2. Tissue-specific variant effect prediction

We next considered combining MMSplice, which models tissue-independent effects together with TSplice, which models differential effects between tissues, to predict the effects associated with genetic variants for any GTEx tissue (Methods). We name this combined model MTSplice. When evaluated on the 51 tissues with at least 10 measured variant effects, MTSplice outperformed MMSplice for 45 out of 51 tissues in terms of root-mean-square error ($P = 1.46 \times 10^{-5}$, paired Wilcoxon test).

2.3. MTSplice predicts brain-specific signals for autism patients

To assess the potential of MTSplice on scoring tissue-specific disease variants, we considered *de novo* mutations that were reported for 1,790 autism spectrum disorder (ASD) simplex families from the Simons Simplex Collection (Yuen et al., 2015) and as provided by Zhou et al (Zhou et al., 2019). The data consists of 127,140 *de novo* mutations, with 65,147 from the proband group and 61,993 from the unaffected siblings. Of those, we further considered the 3,884 mutations lying in exons or in their 300 nt flanking intronic regions and predicted with MMSplice with a delta-logit-PSI greater than 0.05. Overall, MMSplice predicted that variants of the proband group would disrupt splicing more strongly than variants of the control siblings (negative MMSplice scores, Figure 3A, $P = 0.042$, Wilcoxon rank-sum test). The effect was even stronger for the 1,081 loss-of-function (LoF) intolerant genes (Figure 3, $P = 0.0035$, Wilcoxon rank-sum test, Methods). This result is consistent with the report that LoF-intolerant genes are vulnerable to noncoding disruptive mutations in ASD (Zhou et al., 2019) and points to an important contribution of splicing.

We then asked whether MTSplice was able to identify tissue-specific effects of ASD-associated *de novo* mutations. Consistent with the MMSplice results, the *de novo* mutations of the proband group were predicted by MTSplice to more severely disrupt splicing than the *de novo* mutations of the control group for all tissues (Figure 3B). The effect size was larger for the brain tissues (Figure 3B). Since autism is a neurological disorder, these results indicate that MTSplice may be used to prioritize variants that could play a tissue-specific pathogenic role. Besides the brain tissues, the tissues with most pronounced differences were retina, which is also part of the central nervous systems and muscle, which has been associated with autism as well (Paquet et al., 2016). These differences were further amplified when restricting the analysis to the *de novo* mutations in LoF intolerant genes (Figure 3B). Altogether, these analyses demonstrate the value of MTSplice on predicting tissue-specific effects of potentially disease causing mutations.

3. Discussion

We introduced the model MTSplice which quantitatively predicts effects of human genetic variants on RNA splicing in 56 tissues. MTSplice has two components. One component, MMSplice, models constitutive splicing regulatory sequences. The other component, TSplice, models tissue-specific splicing regulatory sequences. The combined model MTSplice outperforms MMSplice on predicting tissue-specific variations in percent spliced-in associated with naturally occurring genetic variants in most tissues of the GTEx dataset. Applying MTSplice to *de novo* mutations from autism spectrum disorder simplex families (Zhou et al., 2019), we found a significantly higher burden for the proband group compared to the control siblings, particularly in brain tissues. These results suggest that MTSplice could be applied for scoring variants with a tissue-specific pathogenic role.

4. Materials and methods

4.1. Dataset

We split the 61,823 cassette exons from ASCOT into a training, a validation, and a test set. The training set consisted of 38,028 exons from chromosome 4, 6, 8, 10-23 and the sex chromosomes. The 11,955 exons from chromosome 1,7,9 were used as the validation set, and the remaining 11,840 exons were used as the test set (chromosomes 2, 3 and 5). Models are evaluated based on their performance on the test set.

4.2. Variant effect estimation

To compute variant effect, we first computed Ψ with MISO for all annotated alternatively spliced exons (MISO annotation v2.0) in all GTEx RNA-Seq samples. Ψ for 4,686 samples from 53 tissues were successfully computed. Second, for each exon, we estimated variant effects using only those samples with a single variant within the exon body and 300 nt flanking of the exon. Third, we estimated the effect associated with the variants as the difference between Ψ averaged across samples homozygous for the alternative allele and Ψ averaged across samples homozygous for the reference allele. We required at least 2 samples in each of these two groups. For simplicity, we did not consider heterozygous samples for estimating the effects because Ψ of heterozygous samples is confounded by allele-specific RNA expression. Also, we did not consider indels.

4.3. The TSplice model

We denote $\Psi_{e,t}$ the percent spliced-in value of the cassette exon e in tissue t. The goal of the multi-tissue splicing model is to predict tissue-specific $\Psi_{e,t}$ from the nucleotide sequence of the given exon S_e. We train the tissue-specific splicing model with multi-task learning, where each task corresponds to a tissue. The model has two input branches. The first input branch consists of the sequence 300 nt upstream of the acceptor and 100 nt downstream of the acceptor (Figure1). In a symmetric fashion, the second input branch consists of the sequence from the donor side, with 100 nt upstream of the donor and 300 nt downstream of the donor. All input sequences are one-hot encoded. The input layer is followed by a 1D convolution layer with 64 filters of length 9. Parameters of the convolution layer are shared by the two input branches, based on the assumption that many sequence motifs are presented both upstream and downstream of the exons. To model the positional dependent effects of splicing motifs, spline transformations (Avsec et al., 2018) are fitted for each of the convolution filters to weight the convolution activations based on the relative input position to donor and acceptor sites. The weighted activations are then concatenated along the sequence dimension.The last fully-connected layer output number of predictions equals the number of tissues (T), corresponding to predictions for each tissue. These are the predictions of the TSplice model mentioned in the rest of the manuscript. During training, logit of the mean Ψ per exon (logit $\overline{\Psi}_e$) was added to these prediction outputs followed by a sigmoid layer. This encourages the model to learn sequence features associated with differential splicing across tissues.

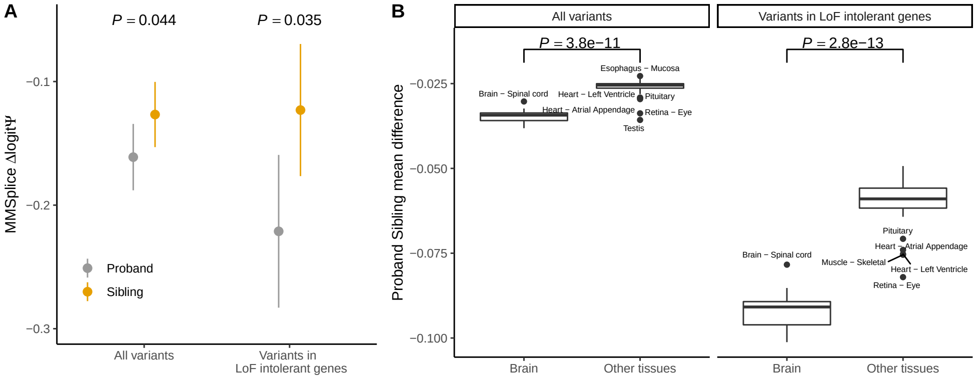

Figure 3. Brain-specific mutational burden on splicing in ASD. **(A)**. Tissue-agnostic variant effect prediction with MMSplice. Splice-region *de novo* mutations (n=3,384, Methods) of the proband group (gray) have significantly lower predicted Δ logit Ψ according to MMSplice compared to those of the unaffected sibling group (orange). The effect size is larger for variants in LoF-intolerant genes (n=1,081). Shown are the means and standard 95% confidence intervals. P-values from one-sided Wilcoxon test. **(B)**. Tissue-specific variant-effect prediction with MTSplice. Distribution of effect size (difference of average Δ logit Ψ for proband versus control siblings *de novo* mutations) for brain tissues (right boxes) and other tissues (left boxes), and for all *de novo* mutations (left panel) or *de novo* mutations in LoF-intolerant genes (right panel) with MTSplice. The predicted effect sizes are more pronounced for brain tissues.

Formally, for each exon, TSplice predicts for each tissue its $\Psi_{\mathrm{e,t}}$ deviation from the mean $\overline{\Psi}_{\mathrm{e}}$ across tissues on logit level. Specifically, we define $\delta\,\mathrm{logit}(\Psi_{e,t})$

$$\delta\,\mathrm{logit}(\Psi_{e,t}) := \mathrm{logit}(\Psi_{e,t}) - \mathrm{logit}(\overline{\Psi}_e) \qquad (1)$$

as the logit Ψ deviation for tissue t and exon e from the logit of mean Ψ across tissues. Denote the number of tissues as T, $\mathrm{logit}(\overline{\Psi}_e) = \mathrm{logit}\left(\frac{1}{T}\sum_{t=1}^{T}\Psi_{e,t}\right)$.

For exon e with input sequence S_e, TSplice predicts the target in $\mathbb{R}^T$: $\mathrm{TSplice}(\mathrm{S}_e) := \Big(\delta\,\mathrm{logit}(\Psi_{e,1}), ..., \delta\,\mathrm{logit}(\Psi_{e,T})\Big)$ corresponding to T tissues.

The tissue-specific $\Psi_{e,t}$ can be predicted with TSplice and the given $\mathrm{logit}(\overline{\Psi}_e)$ computed from the data as:

$$\hat{\Psi}_{e,t} = \sigma\big(\mathrm{TSplice}(\mathrm{S}_e)_t + \mathrm{logit}(\overline{\Psi}_e)\big) \qquad (2)$$

where $\mathrm{TSplice}(\mathrm{S}_e)_t$ is the TSplice predicted $\delta\,\mathrm{logit}(\Psi_{e,t})$. σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$. Note that in eq.1 and elsewhere the average was computed before and not after logit-transformation because it gave more robust results.

4.4. Model training and selection

The model was implemented with keras (version 2.2.4). The Kullback–Leibler (KL) divergence between the predicted and measured Ψ distribution was used as the loss function (3), by considering the percent spliced-in as the probability of the cassette exon to be included in any given transcript.

$$\mathrm{Loss} = \frac{1}{T * E}\sum_{t=1}^{T}\sum_{e=1}^{E}\gamma_{e,t}\Big(\Psi_{e,t}\log(\frac{\Psi_{e,t}}{\hat{\Psi}_{e,t}}) + (1-\Psi_{e,t})\log(\frac{1-\Psi_{e,t}}{1-\hat{\Psi}_{e,t}})\Big), \qquad (3)$$

where

$$\gamma_{e,t} = \begin{cases} 1, & \text{if } \Psi_{e,t} \text{ observed} \\ 0, & \text{otherwise} \end{cases} \qquad (4)$$

Missing values, which typically correspond to tissues in which the gene is not expressed, were masked out in the loss function. Ψ values were clipped to be between [10e-5, 1-10e-5].

4.5. Tissue-specific variant effect prediction

Tissue-specific variant effect $\Delta\Psi_{e,t}$ is predicted as follows (we considered in this study only homozygous cases):

$$\Delta\Psi_{\mathrm{e,t}} = \Psi_{\mathrm{e,t}}^{\mathrm{alt}} - \Psi_{\mathrm{e,t}}^{\mathrm{ref}} \qquad (5)$$

where $\Psi_{\mathrm{e,t}}^{\mathrm{ref}}$ is the measured Ψ for exon e and tissue t with the reference sequence, and $\Psi_{\mathrm{e,t}}^{\mathrm{alt}}$ is the tissue-specific Ψ with the alternative sequence.

The tissue-specific $\Delta\Psi_{\mathrm{e,t}}$ is predicted as follow:

$$\begin{aligned}\Delta\Psi_{\mathrm{e,t}} = &\sigma\Big(\mathrm{logit}(\Psi_{\mathrm{e,average}}^{\mathrm{ref}}) + \mathrm{MMSplice}(\mathrm{S_{ref}}, \mathrm{S_{alt}}) + \mathrm{TSplice}(\mathrm{S_{alt}}, \mathrm{tissue})\Big) \\ &- \Psi_{\mathrm{e,t}}^{\mathrm{ref}}\end{aligned} \qquad (6)$$

5. Acknowledgements

References

Avsec, Ž., Barekatain, M., Cheng, J., and Gagneur, J. Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks, 2018.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J., and Frey, B. J. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

Cheng, J., Nguyen, T. Y. D., Cygan, K. J., Çelik, M. H., Fairbrother, W. G., Avsec, Ž., and Gagneur, J. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.*, 20(1): 48, March 2019.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, 37(9):e67, May 2009.

GTEx Consortium. The Genotype-Tissue expression (GTEx) project. *Nat. Genet.*, 45(6):580–585, June 2013.

Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., Chow, E. D., Kanterakis, E., Gao, H., Kia, A., Batzoglou, S., Sanders, S. J., and Farh, K. K.-H. Predicting splicing from primary sequence with deep learning. *Cell*, 176(3): 535–548.e24, January 2019.

Jian, X., Boerwinkle, E., and Liu, X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.*, 42(22):13534–13544, December 2014.

Ling, J. P., Wilks, C., Charles, R., Leavey, P. J., Ghosh, D., Jiang, L., Santiago, C. P., Pang, B., Venkataraman, A., Clark, B. S., et al. Ascot identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):1–12, 2020.

López-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guigó, R. Are splicing mutations the most frequent cause of hereditary disease?, 2005.

Paquet, A., Olliac, B., Bouvard, M.-P., Golse, B., and Vaivre-Douret, L. The semiology of motor disorders in autism spectrum disorders as highlighted from a standardized Neuro-Psychomotor assessment. *Front. Psychol.*, 7:1292, September 2016.

Parras, A., Anta, H., Santos-Galindo, M., Swarup, V., Elorza, A., Nieto-González, J. L., Picó, S., Hernández, I. H., Díaz-Hernández, J. I., Belloc, E., Rodolosse, A., Parikshak, N. N., Peñagarikano, O., Fernández-Chacón, R., Irimia, M., Navarro, P., Geschwind, D. H., Méndez, R., and Lucas, J. J. Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. *Nature*, 560(7719):441–446, August 2018.

Rosenberg, A. B., Patwardhan, R. P., Shendure, J., and Seelig, G. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):698–711, October 2015.

Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., and Rätsch, G. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007.

Uddin, M., Tammimies, K., Pellecchia, G., Alipanahi, B., Hu, P., Wang, Z., Pinto, D., Lau, L., Nalpathamkalam, T., Marshall, C. R., Blencowe, B. J., Frey, B. J., Merico, D., Yuen, R. K. C., and Scherer, S. W. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder, 2014.

Xiong, H. Y., Barash, Y., and Frey, B. J. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, September 2011.

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., Hua, Y., Gueroussov, S., Najafabadi, H. S., Hughes, T. R., Morris, Q., Barash, Y., Krainer, A. R., Jojic, N., Scherer, S. W., Blencowe, B. J., and Frey, B. J. RNA splicing. the human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, January 2015.

Yeo, G. and Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, 11(2-3):377–394, 2004.

Yuen, R. K. C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., Chrysler, C., Nalpathamkalam, T., Pellecchia, G., Liu, Y., Gazzellone, M. J., D'Abate, L., Deneault, E., Howe, J. L., Liu, R. S. C., Thompson, A., Zarrei, M., Uddin, M., Marshall, C. R., Ring, R. H., Zwaigenbaum, L., Ray, P. N., Weksberg, R., Carter, M. T., Fernandez, B. A., Roberts, W., Szatmari, P., and Scherer, S. W. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.*, 21(2):185–191, February 2015.

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A. K., Yuan, Y., Scheckel, C., Fak, J. J., Funk, J., Yao, K., Tajima, Y., Packer, A., Darnell, R. B., and Troyanskaya, O. G. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.*, 51(6): 973–980, June 2019.